First paper: Statistical rejection sampling improves preference optimization.

1) Full citation: Liu, Tianqi, et al. "Statistical rejection sampling improves preference optimization." arXiv preprint arXiv:2309.06657 (2023).

2) What was the primary problem addressed in the paper?

In the DPO framework, we leverage observed preference pairs to approximate the BT preference. To estimate the optimal policy π^* as a density estimation problem, the optimal way is to fit a policy model on collected preference pairs sampled from π^*. However, DPO uses the collected human preference data from other policies directly in all the experiments and lacks a study on the effect of sampling. Although they propose to sample pairs from the SFT policy and get them labelled by human, it is still not strictly MLE for the preference model due to the mismatch between the sampling distribution and π^*.

3) What was the overall approach?

Statistical rejection sampling is an efficient statistical technique to generate observations from a complex distribution. If we want to generate a distribution of density π^*, we can use π_{sft} as the proposal distribution and follow the general steps of the rejection sampling algorithm.

4) How was the approach validated?

Their experiments use four different approaches to evaluate: Proxy Reward Model, Gold Reward Model, AutoSxS, and Human Evaluation. Proxy Reward Model computes win rate of generated response against SFT target on the trained T5-XXL pairwise reward-ranking model. They also train a PaLM 2-S on the same data as Gold Reward Model. AutoSxS uses PaLM 2-L few-shot in-context learning. Human Evaluation asks human raters to assign a quality score on each response and determine the best one among different systems.

5) What is one thing you particularly liked about the paper?

In reality, it is difficult to obtain human preference pairs directly sampled from the best policy π^* before using DPO to train it. Rejection sampling is a great tool to solve this problem and utilizes the relationship between π^* and π_{sft}, where $\pi_{sft} \propto \pi^*$ up to a deterministic normalization constant Z that is difficult to compute. The approach may solve the problem that DPO tend to overfitting to the preference of the training datasets. This is the thing I particularly like.

6) What is one thing you didn't like about the approach?

The idea of using rejection sampling is great, however, one key step of the algorithm is to estimate the upper bound of the probability ratio π^* / π_{sft} . They just simply sample some responses and calculate the one with the max ratio, which is very naïve and break the condition of the rejection sampling. Advanced methods to find this upper bound is needed.

7) What did you not understand?

As in 6), I don’t understand why they choose this naïve method to estimate the upper bound, which directly affect the acceptance rate and if the sampled data is from the distribution π^*.

Second paper: ARGS: Alignment as Reward-Guided Search

1) Full citation: Khanov, Maxim, Jirayu Burapacheep, and Yixuan Li. "ARGS: Alignment as Reward-Guided Search." arXiv preprint arXiv:2402.01694 (2024).

2) What was the primary problem addressed in the paper?

A pivotal component within RLHF is proximal policy optimization (PPO), which employs an external reward model that mirrors human preferences for its optimization process. However, implementing PPO introduces challenges of unstable and costly training. Furthermore, the need to repeat PPO training when altering the reward model hinders rapid customization to evolving datasets and emerging needs.

3) What was the overall approach?

They steer the decoded outputs of language models in alignment with human preference. At each decoding step, they adjust the model’s probabilistic prediction by a reward signal. This adjustment enables the model to generate text that is not only coherent and contextually relevant but also tailored to satisfy specific alignment criteria or objectives.

4) How was the approach validated?

They leverage three different metrics to evaluate the approach. The first one is average reward, which represents the mean of the reward computed by the reward model across all generations from the test prompts. A higher average reward indicates model continuations that are more closely aligned with the attributes represented in the reward model. The second one is diversity, which aggregates n-gram repetition rates. A higher diversity score indicates the capacity to produce texts with a broad spectrum of vocabulary. The second one is coherence, which is estimated by calculating the cosine similarity between the sentence embeddings of the prompt and its continuation.

5) What is one thing you particularly liked about the paper?

The responses generated by the LLM π^* does not only depend on the policy itself, but also depends on the decoding method used. The RLHF framework generally does not consider and optimize decoding parameters during the training, which cannot guarantee the final generated responses actually satisfied the optimal distribution π^*. One thing I like the most is that this paper consider the decoding optimization.

6) What is one thing you didn't like about the approach?

The idea of considering the reward of each candidate token during the decoding process is great. However, the reward function they are using only consider the current reward of the token, let say, the reward of the current generated tokens. It didn’t consider the expected reward of the full sentences. It is kind of like greedy search. The best token in each step cannot guarantee that it is the best sentence. The reward function needs to be changed to value function, which estimates the expected reward of the full sentences if using the current token to get a global optimal.

7) What did you not understand?

The method in the paper is relatively easy to understand.

Summary:

All two papers are about alignment of the LLM. The first paper improves the DPO algorithm by generating a dataset sampled from the optimal policy π^*. In contrast, the second paper doesn’t use traditional alignment approaches such as DPO or PPO. Indeed, it integrates alignment into the decoding process and enable fast realignment without traditional RLHF framework. It says to have better performance than DPO and PPO.